DES1/AIS1-3L
(Late-News Paper)
Invited

Image Generation with a Unified Generative Adversarial Network Training via Self-Labeling and Self-Attention

Tomoki Watanabe[1] and Paolo Favaro[2]

tomoki8.watanabe@toshiba.co.jp
[1]Toshiba Corporation, Kawasaki, Japan
[2]University of Bern, Bern, Switzerland

Keywords: Image generation, Generative adversarial network, Deep learning, Self-supervised learning

ABSTRACT

Generative Adversarial Network(GAN) is an effective method to obtain an image generation model. We propose a novel GAN training scheme that can handle real images with any level of labeling in a unified manner by introducing a form of artificial labeling. Our scheme consistently improves the quality of generated images.

1 Introduction

Generative Adversarial Networks (GAN) [1] provide an attractive approach to constructing generative models that output samples of a target distribution. In their most basic form, these models consist of two neural networks, a generator G and a discriminator D. The first network is trained to generate samples from some latent representation (typically a sample from a Gaussian distribution), while the second network is trained to distinguish real samples x_r from the generated samples x_f. The most effective GANs seem to benefit greatly from class conditioning. The class information is provided as input to the generator and either injected into the discriminator as an input [2] or through intermediate layers [3] or via a projection [4] or an auxiliary loss [5]. The family of these generators is generically called *conditional GANs* (cGAN).

So far, we have described a GAN training method that exploits the same benefits that conditional GANs enjoy, but without using manually labeled data. When data is partially or fully labeled, it is desirable to take advantage of the available information. Our scheme can seamlessly integrate such available labels and also indirectly transfer their categorical information to the artificial labels. This is possible because our artificial labels are defined relative to the generator and the generator can adapt to a new reference during training.

We evaluate our method on CIFAR-10 [6], STL-10 [7], and SVHN [8] datasets using the BigGAN model [9] and show that our method improves the quality of the generated images in terms of the FID (Fréchet Inception Distance) score [10]. Our method achieves better FID scores than the state-of-the-art GAN and even that of fully supervised cGAN methods.

Our contributions can be summarized as follows:

(1) A unified GAN training that can handle any level of labeling in a unified manner by using: **Self-labeling**: a method to automatically assign labels to real data samples, and **Self-attention**: a method to select real data samples with highly consistent synthetic labels.

(2) Consistent improvement in the FID scores across several datasets (evaluation on CIFAR-10, STL-10, and SVHN).

(3) The ability to outperform class-conditional GANs (fully labeled dataset).

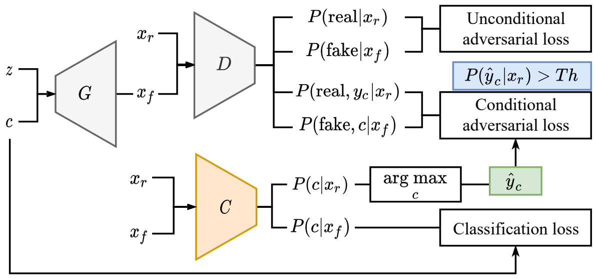


Fig. 1 Architecture of proposed unified GAN training.

2 Method

Our unified GAN training uses a cGAN as backbone, where the discriminator classifies the input into real/fake and image categories. cGANs require semantic labels for training. While the labels of generated data are implicitly defined, the labels of real data are either provided through manual labeling or through our unsupervised self-labeling and self-attention procedures [11]. We show the network architecture of our method in Fig. 1.

2.1 Self-Labeling and Self-Attention

Our objective is to assign artificial labels to unlabeled real images that are used in the conditional adversarial loss. To this purpose, we train a classifier C, which we call *teacher*, on fake images $x_f = G(z, c)$, where the class-correspondence is known. We train the teacher with the cross-entropy loss

$$L_C = H[c, C(A(x_f))]. \tag{1}$$

where H denotes the entropy, the fake image is obtained via $x_f = G(z, c)$ with $z \sim \mathcal{N}(0, I_d)$, c is a random variable (with a discrete Uniform distribution) and also denotes its instance, and A is an image augmentation function. Since the teacher may not be a perfect inverse of G with respect to the conditional label c, we introduce two methods to ensure a high classification accuracy.

First, we use the EMA (Exponential Moving Average) parameters of the teacher $\bar{\theta}_C$ to compute the artificial labels $\hat{y}_c$ of real images, *i.e.*, we compute

$$\hat{y}_c = \arg\max_i C_i(\alpha(x_r); \bar{\theta}_C), \tag{2}$$

where α is a *weak* image augmentation function, *i.e.*, with image transformations close to the identity. Second, because the artificial labels $\hat{y}_c$ are inaccurate especially during the early epochs of the training, we introduce a selection mechanism called self-attention. We first define the *reliability* of the artificial labels via the softmax of the classifier output

$$p_c = \frac{\exp(C_{\hat{y}_c}(\alpha(x_r); \bar{\theta}_C))}{\Sigma_{i=1}^{K} \exp(C_i(\alpha(x_r); \bar{\theta}_C))}, \tag{3}$$

where K is the number of the artificial classes. As we show in the experiments, the reliability yields a high value with real images that are distinctively similar to generated fake images, and when these fake images are well separated into different clusters. Then, self-attention selects real images x_r such that $p_c \geq Th$, where the threshold $Th \in [0, 1]$.

2.2 Training with Artificial (and Real) Labels

The conditional adversarial loss for conventional cGANs uses the supervised class labels y and artificial labels c for real images and fake images respectively as

$$L_D^Y = -E_{x_r,y}[\log P(\text{real}, y|x_r)] - E_{z,c}[\log P(\text{fake}, c|x_f)]. \tag{4}$$

With real images without supervised class labels y, we use instead the artificial labels $\hat{y}_c$.

The discriminator has 2 heads, one for the unconditional fake/real adversarial loss and another for the conditional adversarial loss. The losses for the discriminator L_D and the generator L_G are simply the sum of the corresponding conditional and unconditional losses

$$L_D = L_D^U + L_D^C, \quad L_G = L_G^U + L_G^C. \tag{5}$$

The loss functions

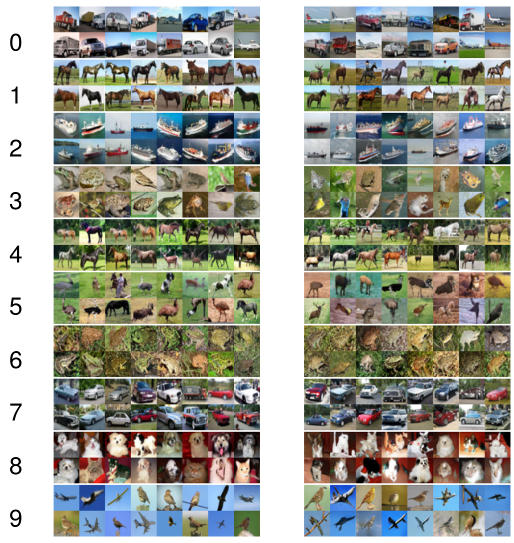
$$L_D^U = -E_{x_r}[\log P(\text{real}|x_r)] - E_z[\log P(\text{fake}|G(z))], \tag{6}$$

$$L_G^U = -E_z[\log P(\text{real}|G(z))], \tag{7}$$

$$L_G^C = -E_{z,c}[\log P(\text{real}, c|G(z, c))], \tag{8}$$

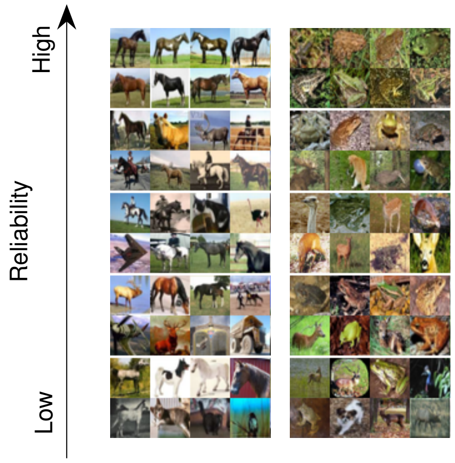
are defined following conventional cGANs. The loss function L_D^C instead is defined so that it can be applied to a dataset with any degree of labeling (from 0% to 100%) as

$$\begin{aligned} L_D^C = & -E_{\{x_r, y|\text{with label}\}}[\log P(\text{real}, y|x_r)] \\ & -E_{\{x_r, \hat{y}_c|\text{no label} \wedge p_c \geq Th\}}[\log P(\text{real}, \hat{y}_c|x_r)] \\ & -E_{x_f, c}[\log P(\text{fake}, c|x_f)]. \end{aligned} \tag{9}$$



(a) Real images (b) Fake images

Fig. 2 Results on the unlabeled CIFAR-10.



(a) Group 1 (b) Group 6

Fig. 3 Examples of the reliability of the artificial labels.

The loss function uses artificial labels $\hat{y}_c$ obtained from the teacher as shown in eq. (2). As explained in subsection 2.1 we calculate the loss only on images where the reliability p_c is higher than a threshold Th, because unreliable labels have an adverse effect on the training of the discriminator. We update the teacher, the discriminator, and the generator simultaneously via eqs. (1) and (5). We can train cGAN on unlabeled dataset, because these loss terms are well-defined even in the absence of real labels.

3 Results and Discussion

We evaluate our method on CIFAR-10, STL-10, and SVHN by using FID scores as a quantitative measure and also visualize samples for a qualitative assessment. The FID scores are computed by using the official implementation [10].

3.1 Unlabeled CIFAR-10

We evaluate the effectiveness of self-labeling and self-attention on CIFAR-10 and summarize the results in Ta-

Table 1 Ablation study on the unlabeled CIFAR-10.

	Self-Labeling	Self-Attention	FID
(A)	-	-	6.96
(B)	✓	-	7.00
(C)	✓	✓	$\mathbf{6.81}$

Table 2 Comparison on the unlabeled CIFAR-10.

Method	FID
BigGAN [9]	14.73
Top-k GAN [12]	13.34
ICR-GAN [13]	9.21
slcGAN [14]	8.95
Top-k ICR-GAN [12]	8.57
Ours	$\mathbf{6.81}$

Table 3 Ablation study on the labeled CIFAR-10.

	Labels	FID
(A)	Artificial	6.81
(B)	Real	4.57
(C)	Artificial & Real	4.35

Table 4 Comparison on the labeled CIFAR-10.

Method	FID
BigGAN [9]	9.06
DiffAug GAN [15]	8.56
MHinge GAN [16]	6.40
FQ-GAN [17]	5.39
Ours	$\mathbf{4.35}$

ble 1. The results show that (B) unreliable artificial labels hurt the performance and (C) refined artificial labels help the training of the GAN compared to using (A) no artificial labels. Our method improves the FID score from 6.96 to 6.81.

In Fig. 2(a), we show real images grouped by their artificial labels, which were learned without supervision. The images are selected via self-attention. By starting from the top, every pair of rows corresponds to one artificial label. We can see that every artificial label identifies images with similar objects, but also that objects across separate labels differ substantially. For example, the groups 0, 2 and 9 contain an object on the ground, in the sea, and in the sky respectively, and the groups 2 and 7 contain ships and cars respectively. In Fig. 2(b), we also show in the same manner images generated using the artificial labels (as input to the generator). Notice the broad diversity and the strong similarity between fake and real images in terms of artificial categories.

To explain the role of the proposed reliability measure p_c (see eq. (3)) for self-attention, we sort real images with the same dominant artificial label based on the magnitude of the reliability. We show in Fig. 3 an evaluation of the consistency between real and artificial labels for two randomly chosen artificial labels. Through visual inspection we find that these labels correspond mostly to the `Horse` (Fig. 3(a)) and `Frog` (Fig. 3(b)) categories. The top and bottom rows correspond to images of high and low reliability respectively. One can observe the higher semantic class consistency (*i.e.*, more `Horse` images in Fig. 3(a) and more `Frog` images in Fig. 3(b)), when the reliability is high.

Finally, we compare our method with the state-of-the-art methods for unsupervised GAN training on CIFAR-10 in Table 2. The methods use different loss functions, but share the same BigGAN generator. Although our generator uses the conditional label input, the basic backbone is the same. Our proposed training shows a significant FID improvement over the previous state-of-the-art (from 8.57 to 6.81).

3.2 Labeled CIFAR-10

As shown in Table 3, our method also improves the FID score when training on labeled datasets. The first column shows the type of labels used in the conditional adversarial loss. We calculate the conditional adversarial loss with either (A) the artificial labels, (B) the real labels, or (C) both of them. The result using both labels yields the best FID. Our method improves the baseline cGAN on the FID from 4.57 to 4.35. The results show that the artificial labels integrate naturally with the real labels and further boost the performance of the generator.

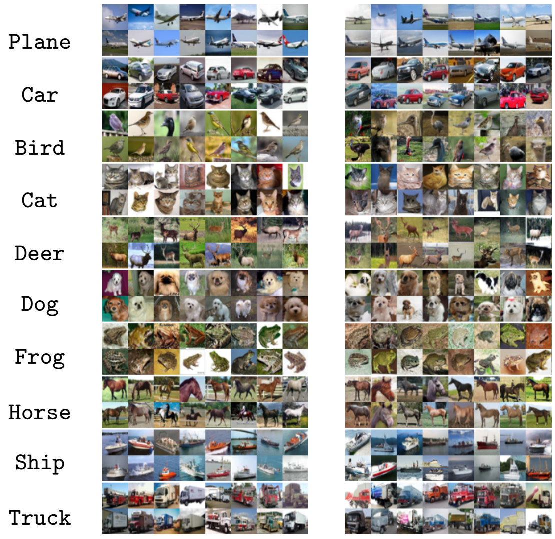
In Table 4, we compare our method with the state-of-the-art cGAN methods on CIFAR-10. As in the unsupervised case, our proposed training shows a significant FID improvement over the previous state-of-the-art (from 5.39 to 4.35).

3.3 STL-10 and SVHN

In Table 5, we compare our method to a baseline without self-labeling and self-attention on STL-10 and SVHN. As we can see from the range of the FID scores, the STL-10 dataset is more complex and the SVHN dataset is simpler than the CIFAR-10 dataset. Another difference is that we use an image size of 48×48 pixels for the experiments on STL-10, while in the CIFAR-10 and SVHN datasets it is of 32×32 pixels. The results show that our method improves the FID score on both datasets compared to the baseline (from 32.85 to 30.91 and from 2.44 to 2.19). In Fig. 5, we show samples of real and generated images on STL-10.

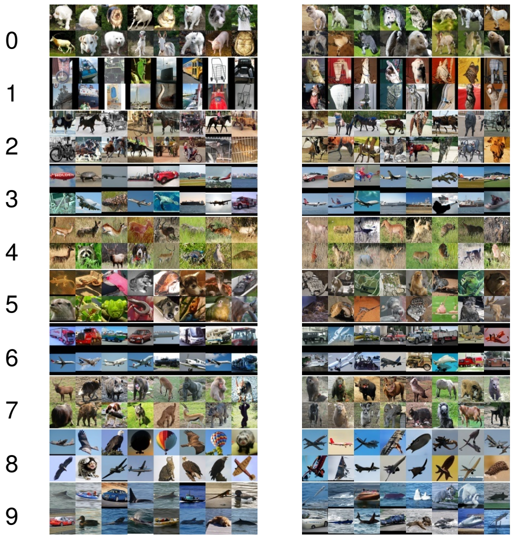
4 Conclusions

We proposed a novel GAN training scheme that can handle different levels of labeling in a unified manner. Our approach is based on using the generator to implicitly define artificial labels and then to train a classifier on purely synthetic data and labels. This classifier can then be used to self-label real data. Its class-consistency is found to correlate well with its classification probability score, which



(a) Real images (b) Fake images

Fig. 4 Results on the labeled CIFAR-10.



(a) Real images (b) Fake images

Fig. 5 Results on the unlabeled STL-10.

Table 5 FID on the unlabeled STL-10 and SVHN..

Dataset	Baseline	+Ours
STL-10	32.85	30.91
SVHN	2.44	2.19

we then use to select samples with a reliable label (self-attention). We evaluated our approach on CIFAR-10, STL-10 and SVHN, and showed that both self-labeling and self-attention consistently improve the quality of generated data.

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proc. NeurIPS 2014*, pp. 2672–2680 (2014).

[2] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *CoRR*, vol. abs/1411.1784 (2014).

[3] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” in *Proc. ICML 2016*, vol. 48, pp. 1060–1069 (2016).

[4] T. Miyato and M. Koyama, “cGANs with projection discriminator,” in *Proc. ICLR 2018* (2018).

[5] A. Odena, C. Olah, and J. Shlens, “Conditional image synthesis with auxiliary classifier GANs,” in *Proc. ICML 2017*, vol. 70, pp. 2642–2651 (2017).

[6] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” in *Technical report* (2009).

[7] A. Coates, A. Y. Ng, and H. Lee, “An analysis of single-layer networks in unsupervised feature learning,” in *Proc. AISTATS 2011*, vol. 15, pp. 215–223 (2011).

[8] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning* (2011).

[9] A. Brock, J. Donahue, and K. Simonyan, “Large scale GAN training for high fidelity natural image synthesis,” in *Proc. ICLR 2019* (2019).

[10] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local nash equilibrium,” in *Proc. NeurIPS 2017*, pp. 6626–6637 (2017).

[11] T. Watanabe and P. Favaro, “A unified generative adversarial network training via self-labeling and self-attention,” in *Proc. ICML 2021*, vol. 139, pp. 11 024–11 034 (2021).

[12] S. Sinha, Z. Zhao, A. Goyal, C. Raffel, and A. Odena, “Top-k training of GANs: Improving GAN performance by throwing away bad samples,” in *Proc. NeurIPS 2020* (2020).

[13] Z. Zhao, S. Singh, H. Lee, Z. Zhang, A. Odena, and H. Zhang, “Improved consistency regularization for GANs,” *CoRR*, vol. abs/2002.04724 (2020).

[14] M. Noroozi, “Self-labeled conditional GANs,” *CoRR*, vol. abs/2012.02162 (2020).

[15] S. Zhao, Z. Liu, J. Lin, J. Zhu, and S. Han, “Differentiable augmentation for data-efficient GAN training,” in *NeurIPS 2020* (2020).

[16] I. Kavalerov, W. Czaja, and R. Chellappa, “cGANs with multi-hinge loss,” *CoRR*, vol. abs/1912.04216 (2019).

[17] Y. Zhao, C. Li, P. Yu, J. Gao, and C. Chen, “Feature quantization improves GAN training,” in *Proc. ICML 2020*, vol. 119, pp. 11 376–11 386 (2020).